BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 9 2011, pages 1343–1344
doi:10.1093/bioinformatics/btr137

Databases and ontologies

Advance Access publication March 12, 2011

UniCarb-DB: a database resource for glycomic discovery

Catherine A. Hayes[1], Niclas G. Karlsson[1], Weston B. Struwe[2], Frederique Lisacek[3], Pauline M. Rudd[2], Nicolle H. Packer[4] and Matthew P. Campbell[4,*]

[1]Department of Medical Biochemistry and Cell Biology, University of Gothenburg, Sweden, [2]National Institute for Bioprocessing Research and Training, University College Dublin, Dublin, Ireland, [3]Proteome Informatics Group, Swiss Institute of Bioinformatics, Geneva, Switzerland and [4]Department of Chemistry and Biomolecular Sciences, Biomolecular Frontiers Research Centre, Macquarie University, Sydney, New South Wales 2109, Australia

Associate Editor: Martin Bishop

ABSTRACT

Summary: Glycosylation is one of the most important post-translational modifications of proteins, known to be involved in pathogen recognition, innate immune response and protection of epithelial membranes. However, when compared to the tools and databases available for the processing of high-throughput proteomic data, the glycomic domain is severely lacking. While tools to assist the analysis of mass spectrometry (MS) and HPLC are continuously improving, there are few resources available to support liquid chromatography (LC)–MS/MS techniques for glycan structure profiling. Here, we present a platform for presenting oligosaccharide structures and fragment data characterized by LC–MS/MS strategies. The database is annotated with high-quality datasets and is designed to extend and reinforce those standards and ontologies developed by existing glycomics databases.

Availability: http://www.unicarb-db.org

Contact: matthew.campbell@mq.edu.au

Received on November 30, 2010; revised on February 14, 2011; accepted on March 9, 2011

1 INTRODUCTION

Glycosylation is a major post-translational modification of proteins that involves the addition of sugar chains to proteins and lipids, a highly complex non-template driven process that generates a diverse glycome. For glycoproteins, the extent of microheterogeneity and macroheterogeneity makes the identification of glycosylation sites and glycan structure determination a challenging analytical problem. Diverse structural modifications can range from a few monosaccharide residues to heavily branched oligosaccharides including modification by non-carbohydrate substituents as well as different linkages and anomeric configurations. Over the past few years, analytical techniques that allow the effective characterization of glycan structures have advanced. In contrast to genomics and proteomics, glycan structures are not template derived and glycomics thus requires well-constructed sets of databases to capture glycan structure and related data that are frequently reported in literature. It is accepted that glycoinformatics is in its infancy but there has been a concerted effort to design glycoinformatic resources to support analytical technologies in the form of a variety of databases, tools and web services (reviewed by Frank and Schloissnig, 2010; Lutteke, 2008). These high-quality databases are capable of providing detailed information about the individual glycans such as structure, biological origin and potential function that are complemented by supporting analytical data. However, to ensure that these resources continue to be relevant it is vital that strides in development and maintenance continue to support the challenges of analytical and structural glycobiology.

*To whom correspondence should be addressed.

2 A GLYCOMICS LC–MS DATABASE

Mass spectrometry (MS) and high-performance liquid chromatography (LC) are established analytical techniques used to address the challenges of glycomics, as they offer high levels of sensitivity and can handle complex mixtures of glycan variations. The coupling of LC with MS systems greatly improves the separation power of MS and strategies have been reported (Ruhaak *et al.*, 2009; Schulz *et al.*, 2002) for both *O*- and *N*-linked carbohydrates. Although glycoinformatic resources have been implemented to assist the interpretation and annotation of data generated by HPLC (Artemenko *et al.,* 2010; Campbell *et al.*, 2008) and MS (Ceroni *et al.*, 2008; Goldberg *et al.,* 2009), there is a distinct lack of informatics infrastructure that supports the growing trend for the LC–MS strategies that are vital to large-scale glycomic and glycoproteomic studies. Here, we present the first release of an LC–MS/MS experimental fragmentation database that will serve as a framework for other glycodatabases. In order to evaluate the format, we have included *O*-linked oligosaccharide MS/MS data from recent publications (Estrella *et al*., 2010; Issa *et al.*, 2010; Karlsson and Thomsson, 2009).

3 DESIGN AND IMPLEMENTATION

The LC–MS/MS database uses the open-source object-relational PostgreSQL database system and was built using the Java environment and the Hibernate object/relational-mapping tool.

3.1 Database features

The design goal was to develop a comprehensive database infrastructure capable of dealing with large volumes of primary information and analytical data. The model provides a platform focused on mass spectrometric data and structural assignments based on fragmentation data that allows users to view all metadata and key experimental data generated from the analysis of glycans. The data model is fully integrated with EUROCarbDB

(von der Lieth *et al.*, 2011) and several new database modules have been developed; whereby the existing definitions have been extended to allow connections between HPLC and MS schemas to support the association of retention time with corresponding MS spectra. These refinements ensure that previously independent HPLC and MS modules are compatible with future developments including a central core for curating sample source information and an overview of all characterized structures.

The information pages for the portal comprise three sections. The first section summarizes the project objectives and a description of the published data stored. For each publication, a listing of all glycan structures can be navigated to retrieve detailed structure and experiment pages. Here, information pertaining to the experiment can be accessed consisting of key attributes to describe the MS instrumentation and settings, HPLC conditions including column type, solvents, gradients and flow rates. The modification and extension of the schemas to connect the two experimental techniques of LC and MS allow for the novel display of retention time and corresponding precursor ion mass that is linked with the MS spectral evidence. The MS spectra section describes any modifications (persubstituted molecules, reducing-end modifications) and ion types in addition to the mode of data acquisition and the MS devices used. In addition, the precursor ion mass for the elucidated structure is supported with an annotated MS/MS fragment peak list. We have focused on providing processed data since the availability of organized and annotated analytical data will facilitate the interpretation of glycomics and glycoproteomics data, and the development of new strategies for the automated, high-throughput identification of glycan structures.

For each entry, a description of the biological source and glycoprotein from which the glycans were isolated is displayed; in addition, a presentation of the glycan type, core type, linkages and anomeric configurations, composition and the supporting MS data can be retrieved. Information content (exact structure, composition or topology) for each structure is obtained from the literature. For example, for a fully annotated structure the linkage information and anomeric configurations must be stated and supported by experimental evidence. If this information is absent or detailed structural profiling was not undertaken, then only the topology of the structure, based on a knowledge of the conserved biosynthetic pathways, is given. Further information detailing the species, tissue and disease state and links to PubMed and UniProt entries are also available.

4 FUTURE PLANS AND CONCLUSIONS

The database outlined here is a novel approach for storing and making available LC–MS glycomics data. Results are clearly reported and all data submitted follows those guidelines proposed and supported by EUROCarbDB. Its availability will serve as a base for the development of new analytical tools for structural data querying and MS spectra interpretation and establishes a link between existing HPLC and MS resources that has been lacking in glycoanalysis. Recording of LC retention time will allow an extra level of confidence in assignment of structures in reproducible LC–MS conditions. By providing a common interface to high-quality data, we believe that this tool can become a vital resource for analyzing glycomic data. At this stage, the pilot study has shown that the database model provides an architecture that adopts standards and practices developed by international consortia, in order to share glycomic experimental data. In addition, several work groups are improving work flows to integrate data processing tools with database storage and querying resources. It is envisaged that methods will be available in the future to allow the capture of raw and processed data that will allow the community to upload published datasets. This capability is vital to the future development of glycomics analysis and it is anticipated that this resource will form the basis for the continued development of glycoinformatic resources.

ACKNOWLEDGEMENTS

We acknowledge the glycobiology groups at Gothenburg University and Macquarie University for testing and feedback.

Funding: Australian Research Council and The Swedish Foundation for International Cooperation in Research and Higher Education.

Conflict of Interest: none declared.

REFERENCES

Artemenko,N.V. *et al.* (2010) GlycoExtractor: a web-based interface for high throughput processing of HPLC-glycan data. *J. Proteome Res.*, **9**, 2037–2041.

Campbell,M.P. *et al.* (2008) GlycoBase and autoGU: tools for HPLC-based glycan analysis. *Bioinformatics*, **24**, 1214–1216.

Ceroni,A. *et al.* (2008) GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J. Proteome Res.*, **7**, 1650–1659.

Estrella,R.P. *et al.* (2010) The glycosylation of human synovial lubricin: implications for its role in inflammation. *Biochem. J.*, **429**, 359–367.

Frank,M. and Schloissnig,S. (2010) Bioinformatics and molecular modeling in glycobiology. *Cell Mol. Life Sci.*, **67**, 2749–2772.

Goldberg,D. *et al.* (2009) Glycan family analysis for deducing N-glycan topology from single MS. *Bioinformatics*, **25**, 365–371.

Issa,S. *et al.* (2010) O-linked oligosaccharides from salivary agglutinin: Helicobacter pylori binding sialyl-Lewis x and Lewis b are terminating moieties on hyperfucosylated oligo-N-acetyllactosamine. *Glycobiology*, **20**, 1046–1057.

Karlsson,N.G. and Thomsson,K.A. (2009) Salivary MUC7 is a major carrier of blood group I type O-linked oligosaccharides serving as the scaffold for sialyl Lewis x. *Glycobiology*, **19**, 288–300.

Lutteke,T. (2008) Web resources for the glycoscientist. *Chembiochem*, **9**, 2155–2160.

Ruhaak,L.R. *et al.* (2009) Oligosaccharide analysis by graphitized carbon liquid chromatography-mass spectrometry. *Anal. Bioanal. Chem.*, **394**, 163–174.

Schulz,B.L. *et al.* (2002) Small-scale analysis of O-linked oligosaccharides from glycoproteins and mucins separated by gel electrophoresis. *Anal. Chem.*, **74**, 6088–6097.

von der Lieth,C.W. *et al.* (2011) EUROCarbDB: an open-access platform for glycoinformatics. *Glycobiology*, **21**, 493–502.